

AD-A074 100

WALTER REED ARMY INST OF RESEARCH WASHINGTON DC
THE EFFECT OF CORRELATION ON THE REPEATED-MEASUREMENTS DESIGN (U)
MAR 61 A LUBIN

F/6 12/1

UNCLASSIFIED

NL

| OF |
ADA
074100



END
DATE
FILMED

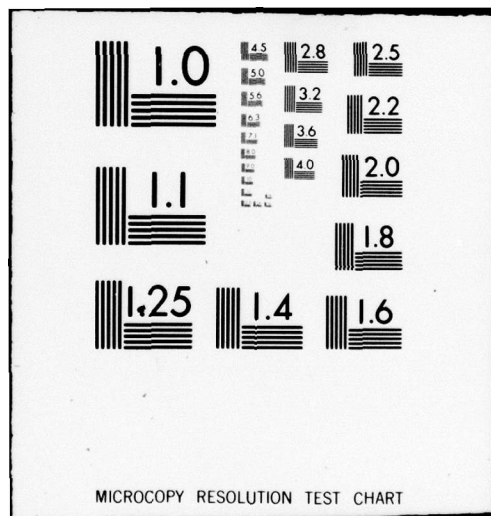
10-79

DDC

END
DATE
FILMED

10-79

DDC



LEVEL II

March 1961

DDC

The effect of correlation on the repeated-measurements design

Ardie/Lubin

SEP 24 1979

Walter Reed Institute of Research

Some difficulties in using the repeated measurements design. The

phrase "repeated measurements design" is used to characterize those experiments where each subject* is tested more than once. Usually this is done to increase the precision of the experiment by eliminating the between-subjects deviance from the estimate of error deviance. Often it is done to avoid multiplying the number of subjects used in the experiment.

The main emphasis of this paper will be on the design where each subject receives only one treatment, applied repeatedly over a period of time, and the chief interest is in the chronic effect of the treatment. An example of such a design would be a drug experiment where each subject is given a constant drug dose every day and tested periodically. Usually each patient is run to a steady state, where the effect of treatment carryover is constant.

In the multiple treatment design, each subject receives a single treatment for a fixed unit of time, but is changed to a different treatment wherever a new time unit starts. A common example would be a drug experiment where a subject might be on drug A the first week, drug B the

+ This paper is a revised version of a talk published in the Proceedings of the conference on the design of experiments, 1960, Office of Ordnance Research, U. S. Army, Box CM, Duke Station, Durham, North Carolina.

* The word subject is used here as a general synonym for the experimental unit of observation.

DISTRIBUTION STATEMENT A
Approved for public release
Distribution Unlimited

U. S. ARMY INFANTRY
HUMAN RESEARCH UNIT

MAY 15 1961

Box 2086

AD A 074100

DDC FILE COPY

79 09 18 083
368 450



DEPARTMENT OF THE ARMY
ARI FIELD UNIT, BENNING

U. S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES
P.O. BOX 2086, FORT BENNING, GEORGIA 31905

PERI-IJ

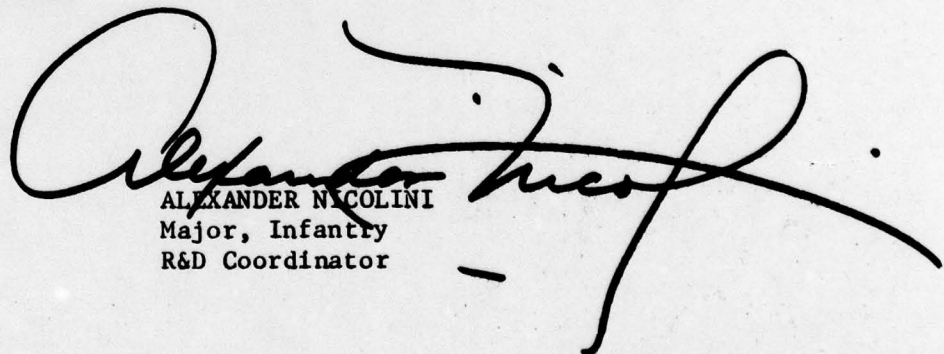
8 August 1979

SUBJECT: Shipment of Documents

Defense Documentation Center
Cameron Station
Alexandria, VA 22314
ATTN: Selection & Cataloging

The Documents in these shipments are approved for public release. The
distribution is unlimited.

FOR THE CHIEF:

A large, stylized handwritten signature in black ink, which appears to read "Alexander Nicolini", is written over the typed name and title.

ALEXANDER NICOLINI
Major, Infantry
R&D Coordinator

Lubin

2.

second week, and so on. The separate effect of each drug is then estimated from the results. Since, in such designs, carryover effects usually are not constant, the solutions suggested in this paper are not appropriate.

The purpose of this paper is to point out that: (a) any repeated measurements on the same organism will in general exhibit statistical dependence; therefore multivariate analysis of variance rather than univariate analysis of variance is appropriate, and (b) multivariate analysis of variance assumes that the carryover effect of a treatment or a test on succeeding treatments is constant and does not depend on the nature of the succeeding treatment, i.e., carryover is additive and does not interact with succeeding treatments.

Most of this paper is concerned with possible experimental and statistical answers to the questions which arise when dependent measures are used in a continuous treatment design. The problem of carryover effects that interact with subsequent treatments is quite different. No answers to this problem are given here; instead we ask if there is, in fact, any way of preserving the advantages of a cross-over design and obtaining unbiased estimates of the treatment effects when carry-over interaction is present.

Let us take a hypothetical psychiatric experiment with a repeated treatment design. Say that a psychiatrist thinks slow reaction times are characteristic of paranoid schizophrenics and he wishes to alleviate this symptom by chronic administration of some tranquillizing drug. He selects a sample of N paranoid schizophrenics, puts each patient on a maintenance dose and starts testing reaction time once a week. At the end of k weeks, and/or

79 09 18 083

| |
|----------|
| For |
| 1 |
| on |
| by Codes |
| special |

the reaction time scores can be arranged as a rectangle, N rows by k columns. The statistical analysis indicated by such texts as Edwards (1950), Lindquist (1953) and McNemar (1949), would be a two-way analysis of variance, with $k-1$ degrees of freedom for the effect of weeks, $N-1$ degrees of freedom for the between-subjects effect, and $(k-1)(N-1)$ degrees of freedom for the subject-by-week interaction effect. Then the significance of the differences between the k weekly means would be assessed by an F ratio using the subject-by-week interaction as the error term. Let us call this ratio the "univariate F ."

One of the basic assumptions for the use of subject-by-week interaction as the error term, is that all observed scores are statistically independent of one another. However, in this hypothetical experiment, it is almost certain that the scores on the first week will have a positive correlation with the scores on the second week, third week, etc.

In 1948, Kogan suggested that if the assumption of independence is not met, the univariate F ratio overestimates the significance of the difference between the k -means. In 1954, G.E.P. Box, in a brilliant article, gave a general technique assessing the effect of departures, from independence and from equal variances, on the univariate F . In general, his conclusions substantiate Kogan's guess; when the null hypothesis is correct and the observations are dependent, the univariate F will exceed the tabled significance levels more often than it should. Roughly speaking — the effect of ^{unequal} correlation between the weeks (i.e., treatments) is to reduce the apparent number of degrees of freedom in the numerator and denominator of the F ratio. Equal correlations leave the d.f. unchanged.

Box's model, and the conclusions he drew, are worth sketching here

since they demonstrate why multivariate analysis of variance, rather than univariate analysis of variance is most generally appropriate for correlated observations. Two assumptions are made:

- a) The vector of scores for any subject is statistically independent of the score vector for any other subject, under the null hypothesis,
- b) Each vector is a sample from the same multivariate normal population.

In terms of our hypothetical psychiatric study, this means that the N paranoids are randomly selected and the relation between the scores of any two weeks, say week s and week t , is bivariate normal. The variance of week t , v_{tt} , need not equal v_{ss} ; v_{et} does not necessarily equal the correlation between any other pair of weeks.

C. R. Rao in 1952 (pp. 239-244) showed how Hotelling's T^2 could be adapted to give an exact test of the differences between correlated means. Basically, Rao takes a linear function of the k scores and compares the mean of this linear function to the variance of the linear function. (A convenient computation routine for this test is given by T. W. Anderson in his 1958 text, par. 5.3.5.)

Using an exact multivariate approach, Box shows that, under the null hypothesis, the true distribution of the univariate F with $(k-1)$ over $(k-1)(N-1)$ d.f. can be approximately represented by the same F value with the degrees of freedom reduced by a fraction, ϵ . This fraction, epsilon, is a function of the k by k covariance matrix.

$$(1) \epsilon = k^2(\bar{v}_{tt} - \bar{v}_{..})^2 / (k-1) \left[\sum_{t=1}^k \sum_{s=1}^k v_{ts}^2 - 2k \sum_{t=1}^k \bar{v}_t^2 + k^2 \bar{v}_{..}^2 \right]$$

where v_{ts} is the covariance of the N pairs of scores from week t and week s , and v_{tt} is the average variance for the k weeks.

The maximum value of epsilon is one, and this is reached only when the k variances are equal and the $\frac{k(k-1)}{2}$ correlations are constant. In this case, Box's approximation gives the exact results; when the correlations are constant and the variances are equal, then the univariate F ratio can be used to give the exact significance level of the differences between the k correlated means.

Geisser and Greenhouse (1958) have shown that the lowest value that epsilon can take is $1/(k-1)$. They argue that since no one has shown what sample estimate of epsilon is most appropriate, and the robustness of epsilon has not been investigated, it is best to use the minimum value of epsilon for a conservative test. This conservative test consists of computing the univariate F , and entering the tabulated F distribution with 1 over $N-1$ d.f. If the result is significant, there is no need to go further; the exact test would be significant. However, if the conservative test is not significant, one can now make an upper-limit test of the univariate F (setting epsilon equal to unity). If an assumed epsilon value of unity gives a non-significant result, then the null hypothesis can be accepted, since no calculated value of epsilon can give a more significant result. However, if using the full degrees of freedom gives a significant result, then the research worker is in a dilemma. Geisser and Greenhouse apparently would next try Box's approximate test, using a sample estimate of epsilon. I would recommend an exact multivariate test such as Rao's. (The sampling distribution of epsilon is not known.)

You can see that the Geisser-Greenhouse approach allows one to bracket the significance level of F with the same amount of computation that is used in the usual two-way analysis of variance. The laborious computations

for an exact multivariate A of V include the data necessary for a two-way A of V. Therefore, it will always be profitable to try the Geisser-Greenhouse approach first, before proceeding to the rest of the distasteful arithmetic necessary for multivariate analysis.

Here it is essential to stop and point out that Box's model explicitly assumes multivariate normality. What alternatives do we have if multivariate normality does not hold or can not be forced by a transformation? As we mentioned previously, the Rao exact multivariate test for differences between correlated means essentially compares the mean of a linear function to the variance of that linear function. The question of multivariate normality can therefore be posed as the question of whether the scores produced by the linear function have a normal distribution. When k is large and correlations are near-zero, we know that the linear function will yield a near-normal distribution of scores. However, if the linear function scores are not normally distributed, the means will have a near-normal shape, assuming the samples of N subjects to be large and selected at random. Therefore the Rao multivariate test will be robust to deviations from normality when N is large or when k is large and the correlations are small.

In those cases where robustness is in question because of small N , high correlation, or other characteristics of the data, it seems to me that the basic strategy should be to resort to the randomization test introduced by R. A. Fisher (1935, par. 21). If we use Box's first assumption, that each subject's vector of scores is independent, and change Box's second assumption to read "each vector is a sample from the same symmetric multivariate distribution," then we will meet Fisher's requirement that the scores for the treatments be drawn from the same population. Since the problem is whether

the means differ significantly, it seems reasonable to use the usual univariate "between treatment means" deviance as the criterion. However, E. S. Pearson (1937) has pointed out that the most powerful criterion depends upon the form of population distribution. For example, when the population distribution is rectangular, midpoints rather than means should be used. The null hypothesis here is that the k scores for any subject are completely interchangeable and any permutation of the k scores can be substituted for the original vector. Since there are N subjects, there are $(k!)^N$ sets of scores. Each set is a possible sample from the original finite set of scores. The between-treatments deviance can be computed for each permutation and we can ascertain where our observed between-treatment deviance falls in the frequency distribution of all possible values from this finite sample. If our observed sample value equals or exceeds the assigned significance level, the means can be judged to be significantly different.

This permutation[†] test preserves one of the advantages of the univariate A of V approach, N can be less than k . (The multivariate methods cannot be applied routinely for N less than k since the inverse of the k by k covariance matrix does not exist.) One disadvantage of the permutation test for differences between means is the requirement that all treatments have identical distribution moments (except for the means). However, the identical distribution assumption apparently is made in every parametric or non-parametric

[†] "Randomization test" is the phrase preferred by most statisticians since the full validity of the test depends upon the way in which the S s were randomly assigned to the treatments. But in our example there is no question of random assignment.

metric statistical test of the difference between two or more samples. The homomorous assumption (of identical distributions) seems to be necessary for generating any statistical test of differences. Some empirical results I have seen suggest that if the distributions are symmetric about their mid-points, they need not be identical; the permutation test is presumably robust to non-identical distributions in these cases.

The basic disadvantage of the permutation test is the extraordinary amount of labor required for even moderate values of N and k .

Suppose, instead of asking if the means are different, we ask if the scores for one week tend to be higher than the scores for other weeks. Then the hypothesis concerns the equality of the rank order averages.

As is well known, Kendall's W , or concordance coefficient, is a simple easily-computed test of this hypothesis. (1948).

Wallis and Friedman independently, and about the same time as Kendall, devised statistics that are algebraically equivalent to Kendall's W .

Essentially, Kendall's W is a permutation test on scores that have been transformed into rankings. The basic assumptions are - score vector independence and identical treatment distributions, exactly the same as those made for Fisher's randomization test—but the laborious computations have disappeared. However, it should be noted that we are now asking a different question - whether the average rank differs significantly between treatments. Does inequality of the average rank imply inequality of the means and vice versa? I have found several empirical examples where Kendall's W was significant but the univariate and multivariate A of V tests fell below significance.

(Generally, one assumes that the rank order statistic and the A of V statistic are testing the same thing, but that the rank-order test is less

powerful. However, the discovery of empirical examples where Kendall's W was significant and the F ratio wasn't, shook my faith in this proposition. Since then, I have learned how to construct examples where the means are exactly identical but the average rank differs significantly. In the construction of these counter-examples, I found it necessary to introduce non-identical distributions, to violate one of the two basic assumptions. However, I understand that it is possible to construct identically skewed distributions where Kendall's W would be more powerful than the F ratio.

Therefore, I would like to raise the explicit question: What are the necessary and sufficient conditions such that rank-order tests are less powerful versions of the analogous A of V tests? This problem transcends the context of repeated measurements. Perhaps situations can be devised such that any rank-order statistic will be more significant than its metric analog. I raise this question - I hope some statistician can answer it.

I am saying that sometimes rank-order tests answer a different question than their metric analogs do. I am not saying that rank-order tests should be abandoned. There may be many occasions when the A of V test is not quite the right way to answer the question - when the major interest is in whether one treatment differs from another treatment, and the amount of the difference is irrelevant. There are other situations where the experimenter, upon reflection, may discover that he is more interested in rank-order than in metric differences.)

Let us now come back to our psychiatric example. You will recall that in our example the psychiatrist had placed his schizophrenic patients on a tranquilizer in the hope that the reaction times would be shortened. Time is a natural unit of measurement and there is little ambiguity there. If he is primarily interested in the therapeutic value of the drug, then the

exact amount of decrease is important. Presumably, any improvement which is insignificant for practical purposes, say a decrease of 1/100 of a second, would be of little therapeutic interest, even if it were statistically significant. However, if his interest is primarily theoretical, for example, he hopes to find whether the delay is at the nerve-muscle junction or is caused by central factors, then any decrease in reaction time will be of interest to him.

Even if he knows that relative and not absolute differences are his main interest, should the psychiatrist use a general test of differences such as Kendall's W, or a test which specifies an a priori rank-order? (for decrease in reaction time should be a monotonic function of number of weeks on the drug). Whenever a set of correlated means has a predicted rank-order, each subject's obtained rank-order can be correlated with the predicted rank-order and the average of all N rank-order correlations can be tested for significance. In 1954 Jonckheere presented an explicit test of this sort, using Kendall's tau. Iyerly (1952) has discussed the distribution of the average Spearman rank-order coefficient, rho. The average rho can be used in place of tau.

Jonckheere's average tau test (as well as the equivalent Spearman form) is unique among non-parametric tests in that there is no parametric analog. So far as I know, there is no exact regression procedure or Hotelling T^2 criterion that can be applied to test for monotonicity. Any metric technique needs a formal specification of the exact mathematical relation between reaction time and weeks, before such a relationship can be tested.*

* Bartholomew (1959) has a metric test of order for uncorrelated means.

This brief survey of the statistical tests appropriate to a continuous treatment design does not, of course, cover all the relevant topics, but it does show there are rational procedures for treating the data which differ considerably from those found in many statistical text-books.

So, to summarize the statistical recommendations in our hypothetical experiment, the psychiatrist might use the Geisser-Greenhouse multivariate A of V approach or Jonckheere's average rank-order coefficient, or he might even decide to compute the laborious but exact test given by Hotelling's T, but he should never make a routine two-way A of V (unless the correlations are equal).

Let me deal briefly with some of the experimental problems raised by repeated measurements. Almost certainly there will be an improvement in reaction time, whether or not the drug is used. The very act of measuring reaction time gives the patient practice on this task, allows him to adjust to the situation, and so on. This quasi-Heisenberg effect is very common with most kinds of repeated measurements. The blood pressure of a subject is usually higher during the first few determinations than on subsequent occasions. The prick of the hypodermic needle can cause significant changes in blood composition until the subject becomes habituated.

One common way of dealing with the problem is to run a control group. This allows us to estimate the trend, without the drug. Another way is to run each patient through the measurement procedure until he reaches a steady state. Control groups are, of course, almost always necessary because of vagaries in the experimental situation, apparatus, etc., but even when controls are used, I advocate running each subject to a steady state. Not only do you usually eliminate any complex trend that may exist, but the intra-subject variation usually decreases markedly. This makes it particularly advantageous to use the intra-subject rather than the inter-subject

variance as error.

But this raises the question of what part of the performance we want to measure. Perhaps it is exactly the factor of learning, habituation, practice, etc., which the experimenter wants to study. In this case, a control group will enable him to assess the effect of a drug on the initial rate of change. In most situations we are interested in the performance of the Subject on a well-learned routine task. When this is, in fact, true, then we may be measuring some factor which is irrelevant to our question when we include measurements taken at a time of rapid learning or habituation.

Let me hasten now to my final point, a sweeping generalized warning against the use of crossover and balanced designs.

If you wish to assess the separate effect of two or more treatments, don't apply the treatment to the same organism. A brief logical justification is as follows: if you're trying to assess the effect of a treatment by itself, then almost certainly you do not have enough previous data to estimate the carryover effect and in particular the interaction of the carryover effect with other treatments. But all standard designs using two or more treatments on the same organism assume that there is no interaction of the carryover effect with preceding or subsequent treatments. (See the 1949 R. L. Solomon design for an exception to this rule.)

Another way of looking at it is to consider the rotation experiment. Here the treatments are applied in predetermined sequence and the problem is the effect of the sequence of treatments on the subject rather than the effects of the individual treatment.

There are countless examples in medicine where the order is all-important, e.g., when weak and strong bacterial strains are injected in an

organism. The enormous difference in the effect of the two rank-orders is the basis for vaccination.

If the experimenter who proposes to use a cross-over design thinks that a rotation experiment with the same treatments would also yield important information, he is assuming that carryover interaction can exist; that treatment A can inhibit or potentiate treatment B. In this case, his estimates of the effect of each treatment from the cross-over design will be hopelessly emmeshed with the carryover interaction effects.

References

- Anderson, T. W. (1958) Introduction to multivariate statistical analysis. New York, Wiley.
- Bartholomew, D. J. (1959) A test of homogeneity for ordered alternatives. Biometrika, 46, 36-48 & 328-335.
- Box, G. E. P. (1954) Some theorems on quadratic forms applied in the study of analysis of variance problems. II. Effects of inequality of variance and correlation between errors in the two-way classification. Ann. Math. Statist., 25, 484-498.
- Edwards, A. L. (1944) Statistical Analysis. New York, Rinehart.
- Fisher, R. A. (1935) Statistical methods for research workers. New York, Stechart.
- Friedman, M. (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J. Am. Statist. Assn., 32, 675.
- Geisser, S. & Greenhouse, S. W. (1958) An extension of Box's results on the use of the F distribution in multivariate analysis. Ann. Math. Statist., 29, 885-891.
- Geisser, S. & Greenhouse, S. W. (1959) On methods in the analysis of profile data. Psychometrika, 24, 95-112.
- Jonckheere, A. R. (1954a) A distribution-free k-sample test against ordered alternatives. Biometrika, 41, 133-145.
- Jonckheere, A. R. (1954b) A test of significance for the relations between m rankings and k ranked categories. Brit. J. Statist. Psychol., 7, 93-100.
- Kendall, M. G. (1938) A new measure of rank correlation. Biometrika, 30, 81.

- Kendall, M. G. (1948) Rank correlation method. London, Charles Griffin & Co., Ltd.
- Kogan, L. S. (1948) Analysis of variance — repeated measurements. Psychol. Bull., 45, 131-143.
- Lindquist, E. F. (1953) Design and analysis of experiments in psychology and education. New York, Houghton Mifflin.
- Iyerly, S. B. (1952) The average Spearman rank correlation coefficient. Psychometrika, 17, 421-428.
- McNemar, Q. (1955) Psychological Statistics. New York, Wiley.
- Pearson, E. S. (1937) Some aspects of the problem of randomisation. Biometrika, 29, 53-64.
- Rao, C. R. (1952) Advanced statistical methods in biometric research. New York, Wiley.
- Solomon, R. L. (1949) An extension of control group design. Psychol. Bull., 46, 137-150.
- Wallis, W. A. (1939) The correlation ratio for ranked data. J. Am. Statist. Assn., 34, 533.

March 1961

Ardie Lubin

THE EFFECT OF CORRELATION OF THE
REPEATED MEASUREMENTS DESIGN